TrustMe: Taking a Critical Approach to LLM-Generated Interpretations of Statistical Analyses.

Sébastien Lê[*] Rémi Mahmoud[†]

Résumé (max 300 mots)

Les grands modèles de langage (LLM) sont de plus en plus utilisés pour produire des interprétations en langage naturel de résultats statistiques, comme en témoigne l'émergence d'outils dédiés dans l'écosystème R (NaileR, EntraineR). Cette évolution ouvre des perspectives prometteuses pour l'enseignement, l'aide à la décision et l'accessibilité des sorties d'analyses. Elle soulève toutefois une question centrale : comment évaluer rigoureusement des interprétations qui peuvent être linguistiquement convaincantes tout en étant scientifiquement inexactes ou insuffisamment justifiées ?

Cette présentation propose une contribution en deux temps, visant à évaluer à la fois la qualité perçue et la cohérence (divergence sémantique, profils d'interprétation), inter-modèles des interprétations générées.

Nous présentons d'abord les résultats d'une enquête auprès d'étudiants amenés à évaluer des interprétations produites par plusieurs LLM à partir d'analyses statistiques courantes (analyse de variance, régression linéaire). Les participants ont jugé chaque réponse selon des critères concrets : structure et clarté générale, ton, exactitude et rigueur statistique, pertinence et synthèse, et utilité.

Cette évaluation centrée utilisateur a mis en évidence des écarts entre qualité rédactionnelle et qualité scientifique, et a motivé la conception d'une méthodologie dédiée à l'analyse critique des résultats générés par des LLM, que nous présentons dans un second temps. Pour un même résultat statistique, nous caractérisons les divergences inter-modèle en plongeant chaque interprétation dans un espace vectoriel (embedding); une matrice de distances est ensuite construite et analysée par multidimensional scaling (MDS) afin de cartographier similitudes et divergences. Nous identifions enfin les deux modèles les plus opposés sur la première dimension et générons, via un LLM, une explication structurée des différences (vocabulaire, ton, style, prudence ou sur-interprétation).

Cette méthodologie a été intégrée dans le package TrustMe afin de comparer systématiquement les interprétations statistiques produites par les LLM.

Mots-clefs (3 à 5) : Statistique - IA - Package - Enseignement

* Institut Agro Rennes-Angers, sebastien.le@institut-agro.fr

† Institut Agro Rennes-Angers, remi.mahmoud@institut-agro.fr